UNIVERSITI TEKNOLOGI MARA

# INTERPRETABLE HYBRID MODELS OF KOLMOGOROV–ARNOLD NETWORKS AND TRANSFORMER FOR MENTAL HEALTH CLASSIFICATION IN LOW-RESOURCE LANGUAGES: A MALAY SOCIAL MEDIA CASE STUDY

**ZAABA BIN AHMAD**

Thesis submitted in fullfillment
of the requirements for the degree of
**Doctor of Philosophy**
**(Computer Science)**

**Faculty of Computer and Mathematical Sciences**

**September 2025**

# ABSTRACT

Depression, anxiety, and stress (DAS) are among the most common global mental health disorders. Social media has become a key outlet where individuals express their psychological states. This research contributes to computational linguistics and mental health informatics by enhancing the classification of DAS in Malay social media, a linguistically diverse and low-resource context marked by extensive colloquial usage. The study addresses several core challenges: the lack of a gold-standard corpus, limitations in existing language models, feature overlap, class imbalance, and issues of model interpretability. A gold-standard annotated corpus is developed using a hybrid strategy that combines expert validation, community a!liations, and self-reported data to ensure reliability and cultural relevance. To address linguistic and computational limitations, this study employs a range of Natural Language Processing (NLP) techniques, including Word2Vec embeddings, Recurrent Neural Networks (RNNs) with attention mechanisms, and transformer-based models such as BERT. To mitigate class imbalance and feature overlap, novel strategies, namely the Class-Aware Attention Model (CAAM) and the Balancing Class Weight Algorithm (BCWA), are introduced, achieving a strong macro average F1-score of 0.88. Further improvement is realised through the integration of Kolmogorov-Arnold Networks (KAN) with BERT. This hybrid KAN-BERT model, enhanced with residual connections, attains a macro average F1-score of 0.92. The structured approach of KAN improves model interpretability by clarifying feature importance, thereby enhancing trust and potential usability in clinical or community mental health settings. Overall, this study delivers a validated corpus, a domain-specific language model, and innovative neural network approaches tailored for low-resource languages. These contributions significantly improve the accuracy and applicability of DAS classification in Malay-language social media, underscoring the role of NLP in addressing mental health challenges in underrepresented linguistic contexts.

# ACKNOWLEDGEMENT

First and foremost, my heart swells with gratitude towards Allah S.W.T., the Almighty, whose boundless blessings have been my constant companion throughout my pursuit of a PhD. With His grace, I have navigated this voyage of discovery and reached the many milestones along the way. The guidance of my research supervisors, Dr Ruhaila Maskat, Professor Dr Hajjah Azlinah Hj. Mohamed, Dr Mike Conway, and Dr Marina Yusoff, has been a beacon in the tumultuous sea of research. I am indebted to them all for their unwavering support and scholarly wisdom. To Dr Ruhaila, in particular, your clear vision and heartfelt sincerity have not only guided me academically but have also deeply inspired me. Working under your mentorship has been one of the great honours of my professional life.

I owe everything to my late parents, Abah (Ahmad Baba) and Mak (

I, whose memories I cherish with every beat of my heart. Your love, prayers, and sacrifices have moulded me into the person I am today. I carry your spirit and teachings with me always. I remain forever indebted for your sacrifices and regret not having had the opportunity to repay them fully. I hope that all my good deeds in this world will accrue to both of you, and that we will meet again in Jannatul Firdaus. My dear wife, Fadhilah, and our beloved children, Iman, Wafiq, Alesha, and Ainan, you are my world. Your love and understanding have been the sanctuary I return to each day. Your prayers have been my shield, and your support the wind beneath my wings. This journey has been ours together, and I am forever grateful for our shared dreams and unbreakable bond. To my siblings and in-laws, your encouragement and prayers have given me strength. In moments of doubt, our family's unwavering belief in me has fortified my resolve.

My heartfelt appreciation extends to the Faculty of Computer and Mathematical Sciences at Universiti Teknologi MARA and the Malaysia Ministry of Higher Education for their support, which has been a cornerstone of my academic journey, providing the resources and environment needed for my research to flourish. Finally, I thank all the individuals who have supported me during this PhD journey, colleagues, friends, and even strangers who have offered their assistance and kindness. This thesis is a testament not just to my efforts but to the collective spirit and goodwill of every individual who has been part of this remarkable chapter of my life.

# TABLE OF CONTENTS

# CHAPTER 1
# INTRODUCTION

Mental health disorders are a pressing global concern, a"ecting millions and burdening healthcare systems and economies. The urgency to address these disorders has intensified, especially following global crises that exacerbate mental health issues. Social media platforms such as Facebook, Twitter, and Reddit have become ubiquitous channels for communication and self-expression, hosting vast amounts of user-generated content that reflect personal experiences and emotions. This rich data provides unprecedented opportunities for researchers to gain insights into public mental health trends and individual psychological states.

Advancements in Natural Language Processing (NLP) enable the analysis of large-scale textual data to classify linguistic patterns associated with mental health conditions. By leveraging NLP techniques, researchers can identify indicators of mental health issues in social media discourse, facilitating timely and informed interventions. However, these techniques predominantly focus on high-resource languages such as English, which benefit from ample linguistic resources and computational models. The Malay language, spoken by millions in Malaysia and neighbouring countries, presents unique challenges for NLP applications. Its linguistic diversity, frequent use of colloquial expressions, and prevalence of code-switching with English classify it as a low-resource language in NLP. The scarcity of tailored linguistic resources and specialised computational models significantly limits nuanced mental health classification in Malay-speaking populations.

This chapter introduces the motivation and objectives of the study while providing a detailed overview of the challenges addressed. It begins by establishing the research background, focusing on the significance of depression, anxiety, and stress (DAS) classification and the challenges posed by Malay-language social media data. The problem statement follows, defining the key gaps in existing methodologies. Research questions and objectives are then outlined to provide a structure for the study. The chapter continues by detailing the scope and highlighting the novel contributions of the research, particularly advancements in NLP and mental health computing, including e"orts to